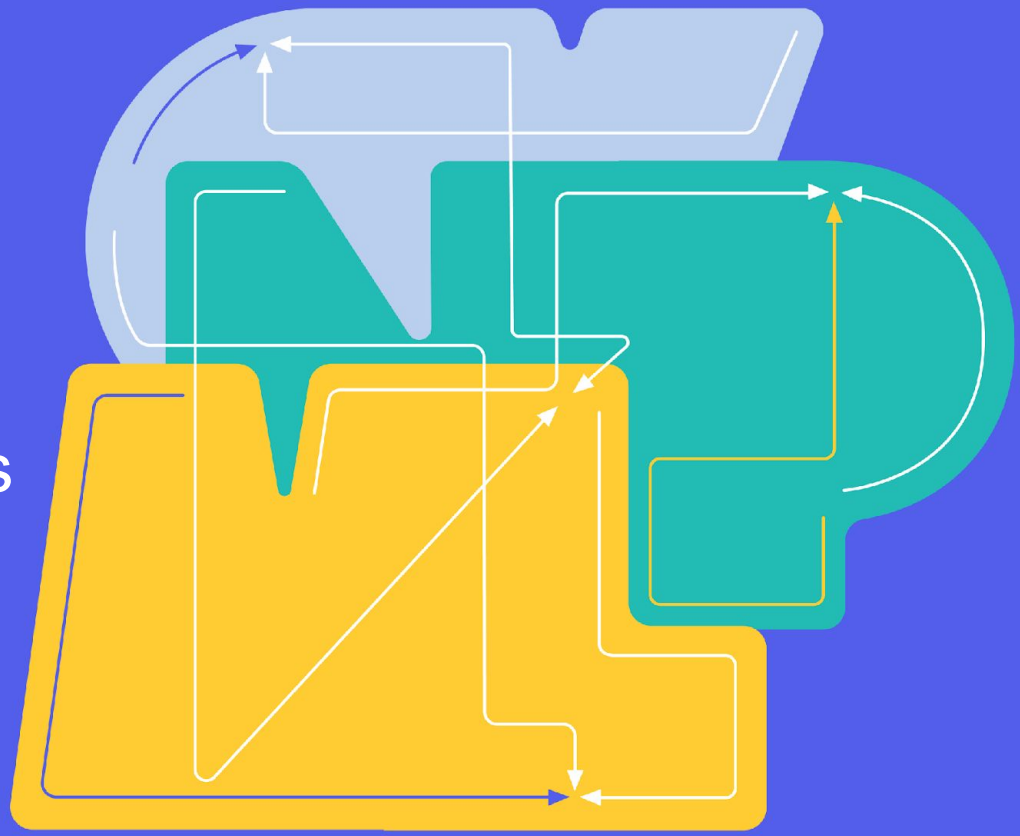
AIRI × Национальный исследовательский Томский государственный университет

31. Ridiculous LLM Compression Techniques



Ignashin Igor, Kiselyov Ivan, Leontyeva Polina, Murzina Anastasiya

Mentor: Bulatov Aydar

Содержание

01

The goal of the project

Ridiculous LLM compression techniques

Modern LLMs are over-parameterized. Research shows up to 50% of layers can be pruned from some models without losing quality [1].

Furthermore, many layers perform near-linear transformations, questioning the need for non-linearities in Transformers [2].

The goal is to explore the possibilities of model compression and to analyze how we can approach it.

[1] Gromov, Andrey, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. “The Unreasonable Ineffectiveness of the Deeper Layers.” arXiv, March 3, 2025. https://doi.org/10.48550/arXiv.2403.17887.
[2] Razzhigaev, Anton, Matvey Mikhalchuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. “Your Transformer Is Secretly Linear.” arXiv, May 19, 2024. http://arxiv.org/abs/2405.12250.

AIRI × Томский государственный университет

02

The results of the experiments

Analyzing heads` similarity

→ **Our approach:**

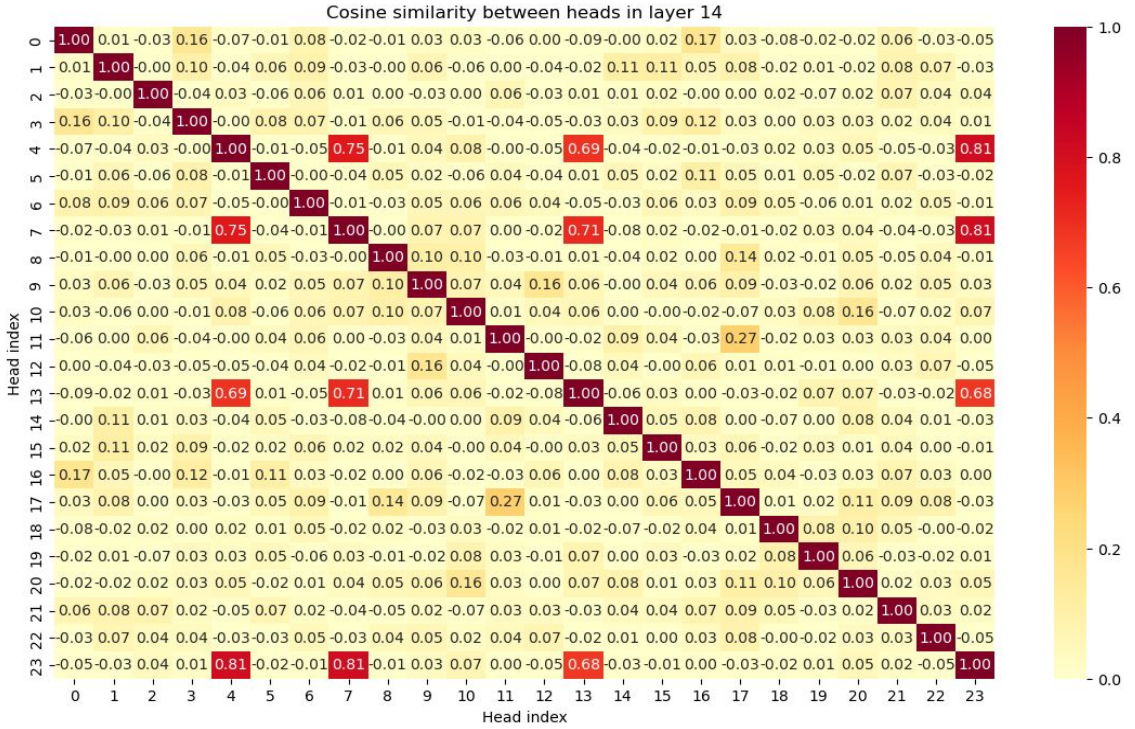
$$\text{similarity} = \frac{x_1 \cdot x_2}{\max(\|x_1\|_2 \cdot \|x_2\|_2, \epsilon)}.$$

→ **Results:**

-Layer 0 - no cos sim

-Intermediate layers > cos sim (L13-14 - 6 pairs)

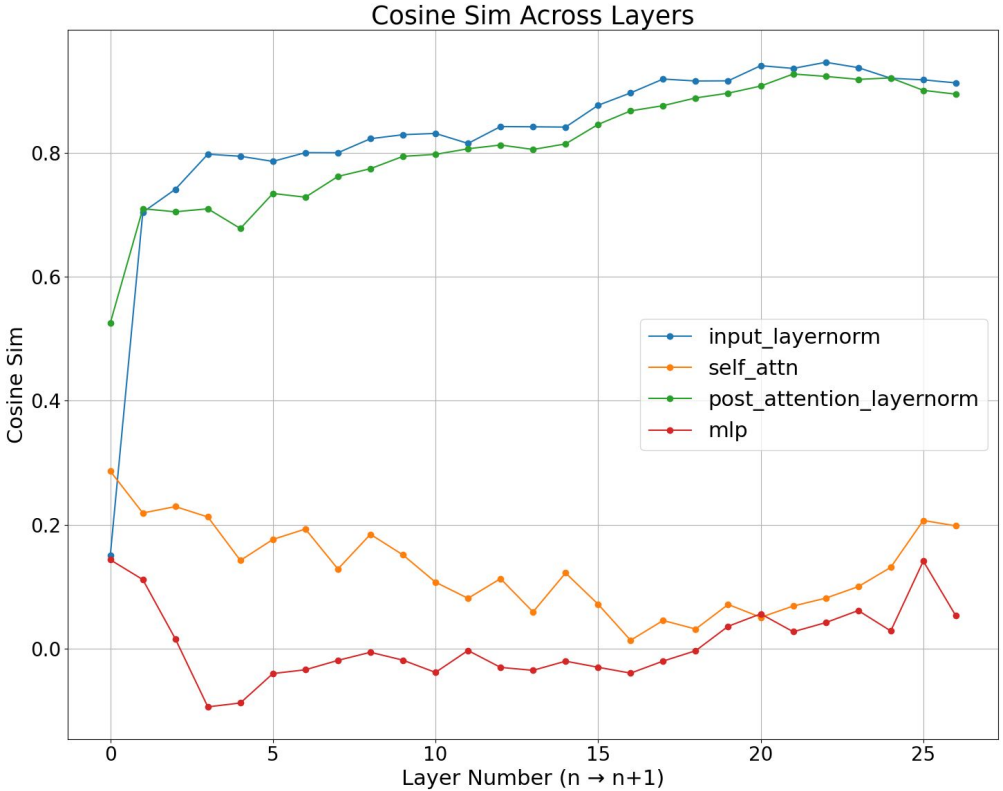
Слой	Найдено пар (cos_sim ≥ 0.5)	Пары голов (сходство)
0	0	-
13	6	4 и 23 (0.778), 7 и 23 (0.746), 13 и 23 (0.729...
14	6	4 и 23 (0.807), 7 и 23 (0.807), 4 и 7 (0.752),...
26	3	7 и 23 (0.744), 4 и 7 (0.622), 4 и 23 (0.590)
27	3	7 и 23 (0.835), 4 и 23 (0.772), 4 и 7 (0.739)



Throw a whole layer ou

- Does something MLP or Attention Layer do? Yes, the hidden state is changed

- What's about the whole layer? Some layers don't change because of the residual connections

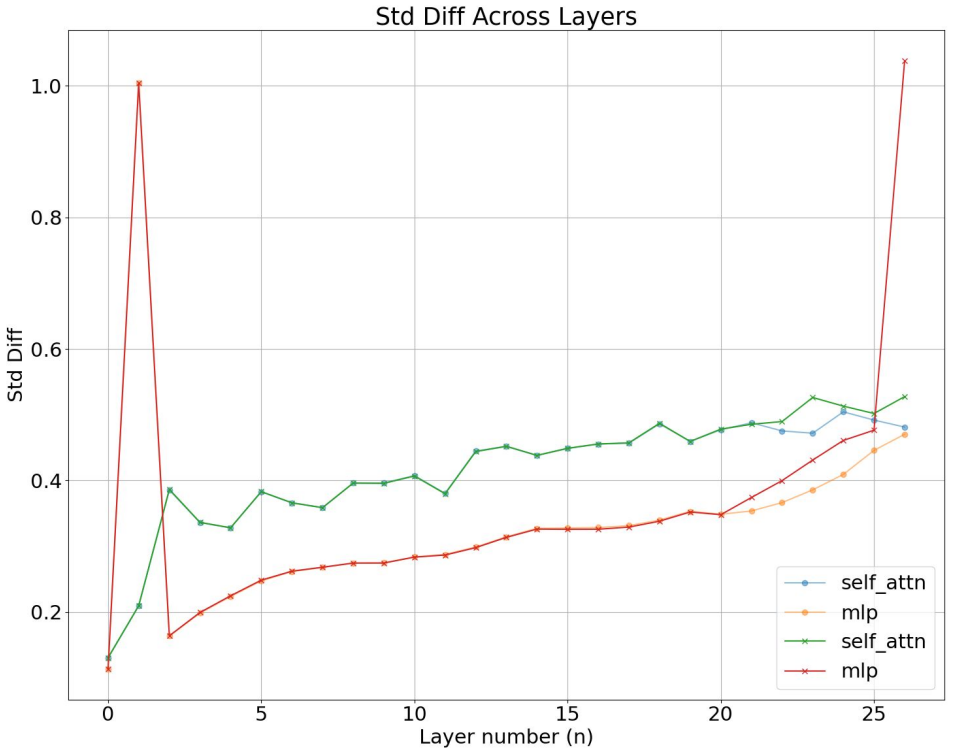


$$\mathbb{E}(\cos\theta^{(l)}) = \frac{1}{N}\sum_{i=1}^{N}\frac{h_i^{(l)}h_i^{(l+1)}}{\|h_i^{(l)}\| \times \|h_i^{(l+1)}\|}$$

AIRI × Томский государственный университет

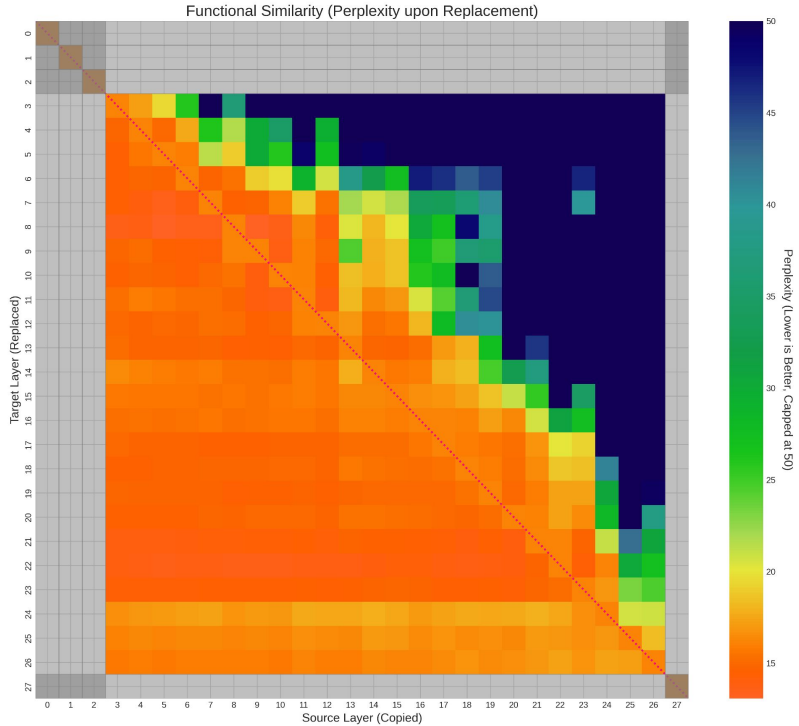
Let’s do it!

- After removing the last MLP layer change the hidden state abnormally strong and heterogenous
- First attention layers are more homogenous than the lasts



$$\sigma^{(l)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{d}\sum_{n=1}^{d}(\Delta h_{i,n}^{(l)} - \mu_n^{(l)})\right)}$$

What if we replace a layer with a copy of another layer?



What do we see and what could it mean

- → Replacing first layers with last layers does not work in general
  - → Layers usually rely on previous computations
- → There are distinct rows and columns
  - → Some layers could be replaced by any layer, some layers could replace any layer
- → The heatmap is symmetrical in neighborhood of x=y
  - → Layers next to each other are often interchangeable

AIRI × Томский государственный университет

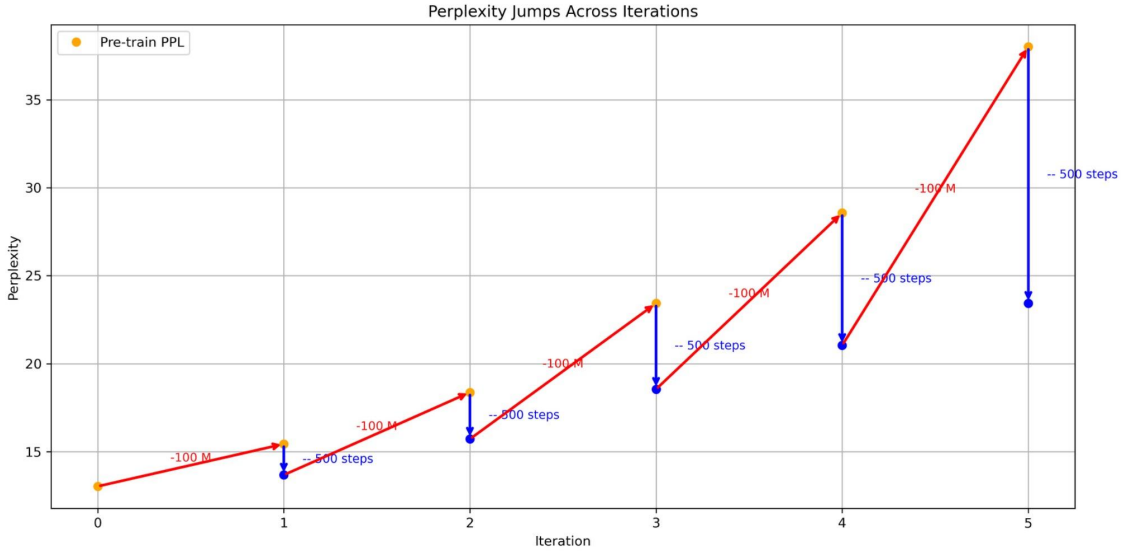
Distillation: Pruning-healing procedure

Iterative LoRA Healing

1) Remove k-th layer
2) Tune LoRA for (k+1)-th
3) Repeat for next layers (5 times)

eval data (1k iters): wikitext-103-raw-v1
train data (0.5k iters): wikitext-2-raw-v1

Setup: LLaMA 3.2 Instruct 3B

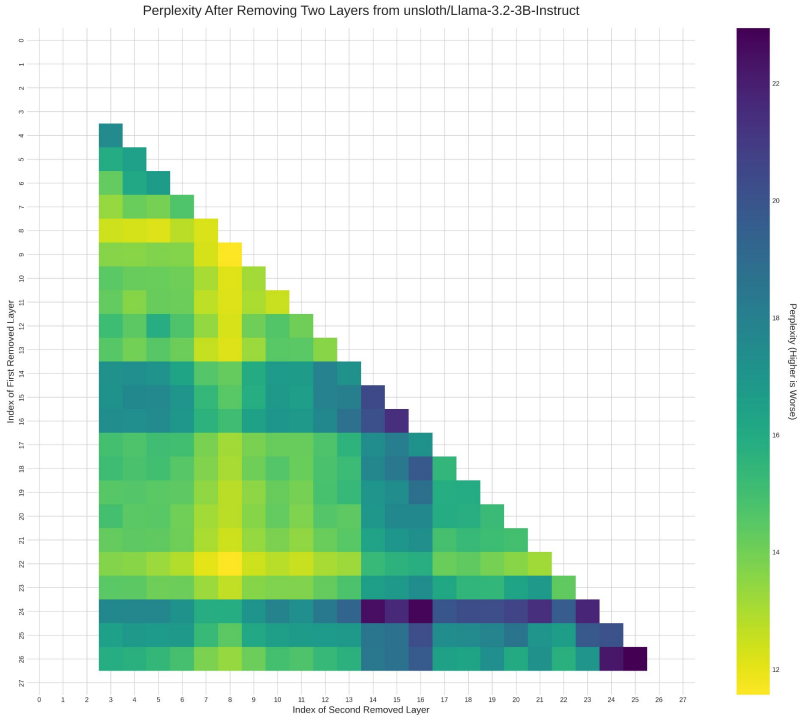


Stage	Perplexity
Initial LLaMA 3.2 3B	12.44
After removing layer	40.36
After LoRA fine-tuning	34.05

- Prune 4 layers and fine-tuning LoRA on the last MLP layer:

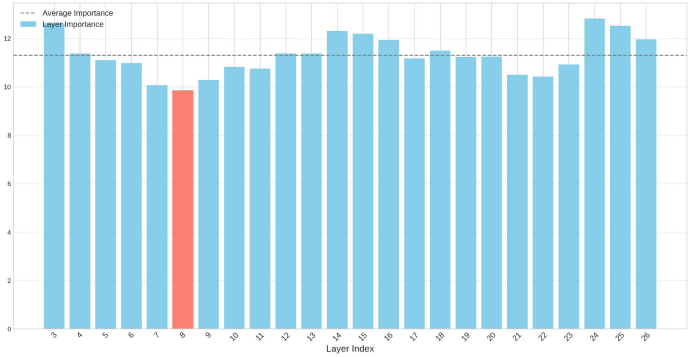
https://github.com/HvostchedUser/Ridiculous-LLM-Compression/tree/main/PruningHealing
https://github.com/ThunderstormXX/PrunningHealing

Layer Importance and Reciprocality



What's that?

- The least impactful layers are 7-13
- Some layers can be removed without significant performance impact while some layers are very important
- Greedy removal of the least impactful layer is suboptimal



AIRI × Томский государственный университет

AIRI

airi.net

airi_research_institute

AIRI Institute

Telegram

AIRI